Asian Journal of Applied Science and Technology (AJAST)
Volume 10, Issue 3, Pages 71-94, July-September 2026

A Comparative Evaluation of Logistic Regression, Random Forest, K-Nearest Neighbors, and Support Vector Machine Algorithms for Machine Learning-Based Lung Cancer Risk Classification Using Accessible Clinical, Demographic, and Lifestyle Risk Factors

Asoshi Paul Anule[1*], Ernest Yakubu Nwuku[2] & Ashraf Ishaq[3]

[1,2,3]*Department of Computer Science, Faculty of Science, Federal University Wukari, PMB 1020, Katsina-Ala Road, Wukari 670101, Taraba State, Nigeria. Corresponding Author (Asoshi Paul Anule) Email: asoshipaulanule@gmail.com*[*]

DOI: https://doi.org/10.38177/ajast.2026.10304

Copyright © 2026 Asoshi Paul Anule et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Article Received: 21 May 2026 *Article Accepted: 27 July 2026* *Article Published: 14 August 2026*

ABSTRACT

Lung cancer remains one of the leading causes of cancer-related mortality worldwide, largely due to late-stage diagnosis and limited access to effective early screening tools. Although low-dose computed tomography (LDCT) has demonstrated potential for early detection, its high cost and limited availability restrict widespread implementation, particularly in resource-constrained settings. Consequently, there is growing interest in computational approaches that utilize clinical, demographic, and biomarker data to support early lung cancer risk identification. This study investigates the effectiveness of machine learning algorithms in predicting lung cancer risk using accessible patient data and blood-related indicators. A machine learning–based classification framework was developed using the publicly available lung_cancer.csv dataset obtained from Kaggle. The dataset contains demographic information, lifestyle factors, smoking history, environmental exposures, and clinical measurements associated with lung cancer risk. Data preprocessing included feature scaling, normalization using the Min–Max method, and feature selection using the SelectKBest technique to identify the most relevant predictors. The dataset was divided into training and testing sets using a 70:30 ratio. Four classification algorithms—Logistic Regression (LR), Random Forest (RF), k-Nearest Neighbors (KNN), and Support Vector Machine (SVM)—were implemented using Python and the Scikit-learn library. Model performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrix analysis. The experimental results indicate strong predictive performance across all evaluated models, with classification accuracies exceeding 93%. Among the algorithms, the Random Forest model achieved the highest accuracy of 95.0%, along with high precision, recall, and F1-score values. Feature selection identified several influential predictors, including smoking exposure, air pollution index, pulmonary function indicators, and inflammatory biomarkers, highlighting their relevance in lung cancer risk prediction.

Keywords: Lung Cancer; Machine Learning; Risk Classification; Logistic Regression; Random Forest; K-Nearest Neighbors; Support Vector Machine; Blood-Based Biomarkers; Feature Selection; Clinical Risk Factors; Lifestyle Risk Factors; Predictive Modeling.

1.0. Introduction

Lung cancer continues to be one of the most serious health challenges worldwide and remains a leading cause of cancer-related mortality. According to the World Health Organization, the disease is responsible for roughly 1.8 million deaths each year, highlighting its major global health burden. Data from the International Agency for Research on Cancer indicate that approximately 2.2 million new cases were diagnosed globally in 2020, ranking lung cancer as the second most frequently diagnosed malignancy after breast cancer (IARC, 2023).

Clinically, lung cancer is broadly divided into two major types. Non-small cell lung cancer (NSCLC) accounts for the majority of cases around 85% while the remaining proportion consists of small cell lung cancer (SCLC) (Sharma, 2022). These categories differ considerably in terms of growth patterns, treatment approaches, and prognosis.

Cigarette smoking remains the most significant risk factor, contributing to nearly 85% of lung cancer cases. Nevertheless, other determinants such as environmental pollution, occupational exposure to carcinogens, and genetic susceptibility also play important roles in disease development (WHO, 2023). Although several high-income countries have observed declining incidence rates due to sustained anti-smoking policies and

improved awareness, the opposite trend is occurring in many low- and middle-income regions where tobacco use and environmental risks are increasing (Sharma, 2022).

Early identification of lung cancer dramatically improves patient outcomes. When NSCLC is diagnosed at an early stage, the five-year survival rate may reach approximately 70%. However, survival drops to below 10% once the disease progresses to advanced stages. A major challenge is that early lung cancer rarely produces clear symptoms, leading to delayed diagnosis in many patients (WHO, 2023).

Screening with low-dose computed tomography (LDCT) has proven effective in reducing lung cancer mortality and is recommended for individuals at high risk. Studies indicate that LDCT screening can reduce lung-cancer-related deaths by roughly 20%. Despite this benefit, widespread implementation remains limited due to the cost of imaging infrastructure, the requirement for specialized expertise, and relatively low participation rates among eligible individuals (Schwartz et al., 2025). Participation rates in some screening programs remain below 50%, particularly in settings with limited healthcare resources.

Because of these limitations, researchers have increasingly focused on non-invasive diagnostic strategies, particularly blood-based biomarkers. Such approaches offer a potentially accessible and cost-efficient method for early detection. Biomarkers present in circulating blood include proteins, cell-free DNA fragments, microRNAs, and metabolic products that may reflect tumor activity (Arguís et al., 2024). In addition, routine clinical blood parameters such as white blood cell count, red blood cell distribution width, and serum albumin levels are being explored as possible indicators of lung cancer risk because they are already widely available in standard laboratory testing (Wu et al., 2019).

The growing availability of biomedical data has encouraged the application of machine learning (ML) techniques in cancer detection. ML algorithms can identify subtle patterns within complex datasets that may not be easily recognized through conventional statistical approaches (Schwartz et al., 2025). Previous investigations have demonstrated promising results when ML models are trained using routine blood test variables. For instance, Wu et al. (2019) developed a model using 19 blood parameters that achieved sensitivity of 96.3% and specificity of 95.0%. Similarly, Schwartz et al. (2025) reported that an ML system based on 22 routine blood markers could estimate lung cancer risk several years prior to diagnosis with an accuracy exceeding 70%.

Although significant advances have been made in lung cancer detection through machine learning applications involving circulating tumor DNA (ctDNA), microRNAs, methylation biomarkers, proteomics, metabolomics, and multi-omics data integration, many of these approaches require expensive laboratory infrastructure, specialized sequencing technologies, and advanced clinical expertise. Consequently, their adoption remains challenging in low-resource healthcare environments.

Furthermore, existing studies often focus on developing highly specialized predictive models using complex molecular datasets, while comparatively limited attention has been given to evaluating the effectiveness of conventional machine learning algorithms using readily available clinical and lifestyle-related risk factors. Many healthcare institutions routinely collect demographic information, smoking history, environmental exposure

indicators, and respiratory health measurements that could potentially support early risk assessment without requiring sophisticated biomarker testing.

In addition, there remains insufficient comparative evidence regarding the relative performance of widely used machine learning algorithms such as Logistic Regression, Random Forest, Support Vector Machine, and K-Nearest Neighbors when applied to accessible lung cancer risk prediction datasets. Identifying the most effective algorithm for such practical healthcare scenarios could facilitate the development of affordable and scalable decision-support tools.

Therefore, this study conducts a comparative evaluation of four machine learning algorithms for lung cancer risk classification using clinical and lifestyle risk factors, with the objective of identifying the most effective model for supporting early risk assessment in resource-constrained healthcare settings.

1.1. Study Objectives

The specific objectives of this study are to:

1) Review and synthesize existing literature on machine learning applications for lung cancer detection, including blood-based, molecular, and clinical/lifestyle-based approaches.

2) Develop a machine learning-based classification framework using the publicly available lung_cancer.csv dataset to estimate lung cancer risk from clinical and lifestyle variables.

3) Apply data preprocessing techniques, including Min-Max normalization and SelectKBest feature selection, to identify the most relevant predictors of lung cancer risk.

4) Implement and compare four supervised classification algorithms, namely Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM), under identical experimental conditions.

5) Evaluate and compare the performance of the four algorithms using accuracy, precision, recall, F1-score, and confusion matrix analysis.

6) Identify the most influential clinical and lifestyle predictors of lung cancer risk and discuss the practical implications of the findings for affordable, scalable risk screening in resource-constrained healthcare settings.

2.0. Literature Review

Recent research efforts have focused on addressing the limitations of current lung cancer screening techniques through the integration of machine learning with emerging biomedical data sources. One prominent direction involves the use of multimodal data integration, where different biological signals are combined to enhance diagnostic performance. Studies suggest that integrating circulating tumor DNA (ctDNA) with other molecular markers such as exosomal RNA, proteomic profiles, or metabolomic signatures can significantly improve detection accuracy.

For example, Thalambedu et al. (2025) reported that combining ctDNA fragmentomic features with protein biomarkers resulted in a sensitivity of 86.4% for early-stage NSCLC detection. Large-scale initiatives, including

CancerSEEK and the Circulating Cell-Free Genome Atlas (CCGA), have also demonstrated the effectiveness of multi-analyte models that merge genomic mutations, DNA methylation patterns, and protein biomarkers. These studies suggest that integrating multiple biological signals through machine learning could provide more reliable diagnostic models than approaches based on a single biomarker category.

2.1. Machine Learning in Lung Cancer Detection

Machine learning has emerged as a promising approach for improving lung cancer detection by identifying complex relationships among clinical, demographic, and biological variables that may not be evident through traditional statistical techniques. Early studies demonstrated that routine healthcare data could be effectively utilized for risk prediction without relying solely on expensive imaging or molecular diagnostics. For example, Wu et al. (2019) developed a machine learning framework based on routine blood indices and reported excellent diagnostic performance, achieving sensitivity and specificity values exceeding 95%. Their findings highlighted the feasibility of using readily available laboratory data as a non-invasive tool for lung cancer identification.

Building on this foundation, Dritsas and Trigka (2022) evaluated several machine learning algorithms for lung cancer risk prediction and demonstrated that ensemble-based approaches generally outperformed conventional statistical models. Their study emphasized the importance of feature engineering and algorithm selection in improving classification accuracy, while also illustrating the growing role of machine learning in supporting clinical decision-making.

More recently, Gandhi et al. (2023) provided a comprehensive assessment of artificial intelligence applications in lung cancer management. Their review showed that machine learning techniques are increasingly being integrated across the cancer care continuum, including early detection, diagnosis, treatment planning, and prognosis prediction. Although these approaches have demonstrated considerable potential for enhancing patient outcomes, the authors noted that issues related to data quality, model transparency, and clinical validation remain significant barriers to widespread implementation.

Collectively, these studies demonstrate that machine learning can substantially improve lung cancer detection accuracy. However, much of the existing work focuses either on highly specialized datasets or advanced computational frameworks, creating a need for further comparative investigations using accessible clinical and lifestyle-related risk factors that are routinely collected in healthcare settings.

2.2. Blood-Based Biomarkers and Molecular Diagnostics

The limitations associated with imaging-based screening have stimulated extensive research into blood-based biomarkers as non-invasive alternatives for early lung cancer detection. Advances in liquid biopsy technologies have enabled the identification of circulating molecular signatures that can reveal the presence of malignancy before clinical symptoms become apparent.

One important area of investigation involves cell-free DNA (cfDNA) analysis. Mathios et al. (2021) demonstrated that fragmentation patterns within circulating DNA could serve as highly accurate indicators of lung cancer. Their

fragmentome-based approach achieved an area under the curve (AUC) of approximately 0.98, illustrating the diagnostic value of tumor-derived genomic alterations detected through blood samples.

Similarly, Hu et al. (2023) explored the diagnostic utility of DNA methylation biomarkers and developed a seven-marker methylation panel capable of distinguishing lung cancer patients from non-cancer individuals with an AUC of 0.94. The study highlighted the growing importance of epigenetic signatures as reliable indicators of early tumor development.

Beyond genomic and epigenomic markers, researchers have increasingly investigated proteomic and metabolomic biomarkers. Chen et al. (2023) integrated blood protein and metabolite profiles into a predictive model that achieved high sensitivity and specificity for non-small cell lung cancer detection. Their findings suggest that combining multiple biomarker categories can improve diagnostic performance compared to single-marker approaches.

Likewise, Davies et al. (2023) employed plasma proteomic profiling to identify biomarkers associated with lung cancer risk several years before clinical diagnosis. Their machine learning model demonstrated strong predictive capability and underscored the potential of blood-based biomarkers for population-level screening and risk stratification.

Complementary lines of research have further broadened this biomarker landscape, including microRNA-based diagnostic panels (Fehlmann et al., 2020; Kim et al., 2023; Poh et al., 2023; Røe et al., 2024), circulating exosomal protein biomarkers (Liu et al., 2024), untargeted serum metabolomic profiling (Ji et al., 2023), peripheral blood transcriptomic signatures (Li et al., 2025), and cfDNA-based liquid biopsy frameworks (Heitzer et al., 2023). Broader reviews have also synthesized this body of evidence (Maharjan et al., 2020), while related work has examined blood-based biomarkers in the context of treatment response, such as immune checkpoint blockade (Tsai et al., 2024). Taken together, these studies demonstrate that blood-based molecular diagnostics offer promising opportunities for early lung cancer detection. Nevertheless, the specialized laboratory infrastructure, sequencing technologies, and analytical expertise required for many of these approaches may limit their adoption in resource-constrained healthcare environments.

2.3. Multi-Omics and AI-Based Detection

Recognizing the limitations of single-biomarker strategies, recent research has increasingly focused on integrating multiple biological data sources through artificial intelligence and machine learning techniques. Multi-omics approaches combine genomic, transcriptomic, proteomic, metabolomic, and epigenomic information to provide a more comprehensive representation of disease biology.

Kwon et al. (2023) demonstrated the effectiveness of this strategy by integrating liquid biopsy-derived biomarkers, including circulating tumor DNA content, tumor markers, and copy number variations, within a machine learning framework. Their findings showed that multi-omics models consistently outperformed approaches based on individual biomarker categories, suggesting that complementary biological signals can enhance diagnostic accuracy.

Similarly, Thalambedu et al. (2025) highlighted the growing convergence of artificial intelligence and circulating tumor DNA analysis for non-small cell lung cancer detection. The authors reported that AI-driven integration of ctDNA fragmentomic characteristics with protein biomarkers significantly improved early-stage detection performance. Their review further emphasized the role of advanced machine learning algorithms in extracting clinically meaningful patterns from large and heterogeneous molecular datasets.

Although multi-omics approaches have achieved impressive diagnostic results, several challenges remain. Data integration complexity, high implementation costs, limited standardization, and the need for extensive computational resources continue to hinder routine clinical adoption. Consequently, there remains a need for alternative predictive frameworks that can deliver reliable performance using more accessible clinical and lifestyle-related variables.

2.4. Explainable AI and Future Directions

As machine learning models become increasingly sophisticated, concerns regarding interpretability and clinical trust have gained greater attention. Healthcare professionals often require transparent explanations of model predictions before incorporating AI-generated recommendations into clinical practice. Consequently, explainable artificial intelligence (XAI) has emerged as a critical research area within medical machine learning.

Flyckt et al. (2024) demonstrated the practical value of explainable AI by developing a lung cancer prediction model based on routine blood test results and smoking history. Their approach incorporated interpretable machine learning techniques that enabled clinicians to understand the relative importance of specific predictors contributing to risk assessments. The study illustrated how transparency can facilitate clinical acceptance while maintaining strong predictive performance.

Beyond model interpretability, future research is increasingly focused on addressing challenges related to data privacy, scalability, and collaborative model development. Ankolekar et al. (2025) examined the application of federated learning in cancer research and reported that decentralized machine learning approaches often achieved superior generalization performance compared with conventional centralized models. By allowing institutions to train models collaboratively without sharing sensitive patient data, federated learning offers a promising pathway toward the development of robust and globally representative cancer prediction systems.

Overall, future advances in lung cancer detection are expected to be driven by the integration of explainable AI, federated learning, automated machine learning, and multi-modal data analytics. These innovations have the potential to improve diagnostic accuracy while simultaneously enhancing transparency, trustworthiness, and clinical applicability, thereby supporting the broader adoption of AI-enabled decision-support systems in oncology.

2.5. Literature Synthesis

The reviewed literature demonstrates substantial progress in machine learning-assisted lung cancer detection. However, most studies rely on advanced molecular biomarkers, specialized laboratory procedures, or expensive sequencing technologies. While these approaches often achieve excellent predictive performance, their practical deployment remains challenging in many healthcare environments. Furthermore, relatively few studies have

systematically compared conventional machine learning algorithms using readily available clinical and lifestyle risk factors. This limitation highlights the need for practical and scalable predictive frameworks that can support early lung cancer risk assessment without dependence on complex molecular diagnostics.

Overall, future research is expected to focus on developing large-scale, multi-parameter models that integrate diverse biological and clinical data sources. Improvements in sequencing accuracy, enhanced liquid biopsy technologies, and collaborative federated learning frameworks will likely play key roles in advancing reliable early detection systems. If these challenges are successfully addressed, machine learning based blood tests could become a practical and scalable solution for lung cancer screening and early diagnosis.

This study contributes to the existing body of knowledge in four major ways:

1. It evaluates the predictive capability of commonly used machine learning algorithms for lung cancer risk classification using readily available clinical and lifestyle risk factors.

2. It provides a comparative assessment of Logistic Regression, Random Forest, Support Vector Machine, and K-Nearest Neighbors under identical experimental conditions.

3. It identifies the most influential risk factors contributing to lung cancer risk prediction through feature selection techniques.

4. It provides evidence regarding the suitability of traditional machine learning approaches for supporting lung cancer risk assessment in healthcare environments where advanced molecular diagnostic technologies may not be readily available.

3.0. Methodology

3.1. Introduction

This section presents a comprehensive explanation of the research methodology utilized. Furthermore, it thoroughly discusses all the experimental method employed to implement the proposed Machine Learning Classification of Lung Cancer Risk Using Clinical and Lifestyle Risk Factors. It also provides an overview of the dataset used, the techniques applied for data preprocessing and feature selection, a description of the machine learning algorithms, and, most importantly, the performance evaluation metrics used to validate the performance of each model.

3.2. Research Methodology Overview

In any research undertaking, it is important to adopt techniques grounded in sound engineering principles to ensure the efficient development and evaluation of the project's feasibility. A methodology refers to a systematic strategy or set of techniques employed to analyze and present the desired outcome. Building a machine learning model for lung cancer risk classification using clinical and lifestyle risk factors is challenging and requires an effective methodology. Hence, to construct the envisioned machine learning model, this research has implemented a methodology approach consisting of three distinct phases. The initial stage involves accessing the dataset through the use of the pandas library, which offers many approaches for reading datasets in different file formats. In this

study, the dataset format chosen is comma-separated values (CSV). The second phase of the procedure involves data preparation, which aims to carry out feature extraction. In the data preprocessing stage, Min-Max normalization was applied to numerical variables, while SelectKBest with ANOVA F-score (f_classif) was employed for feature selection to standardize and encode the dataset's labels and features. This was performed before the selection of pertinent features through the examination of the correlation matrix of the original dataset. In the final phase, the dataset is trained using Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). Before this, the dataset is divided into a 70:30 ratio for training and testing purposes. The training data is utilized to train the four supervised learning classification models, while the test data is employed to assess the accuracy of the model.

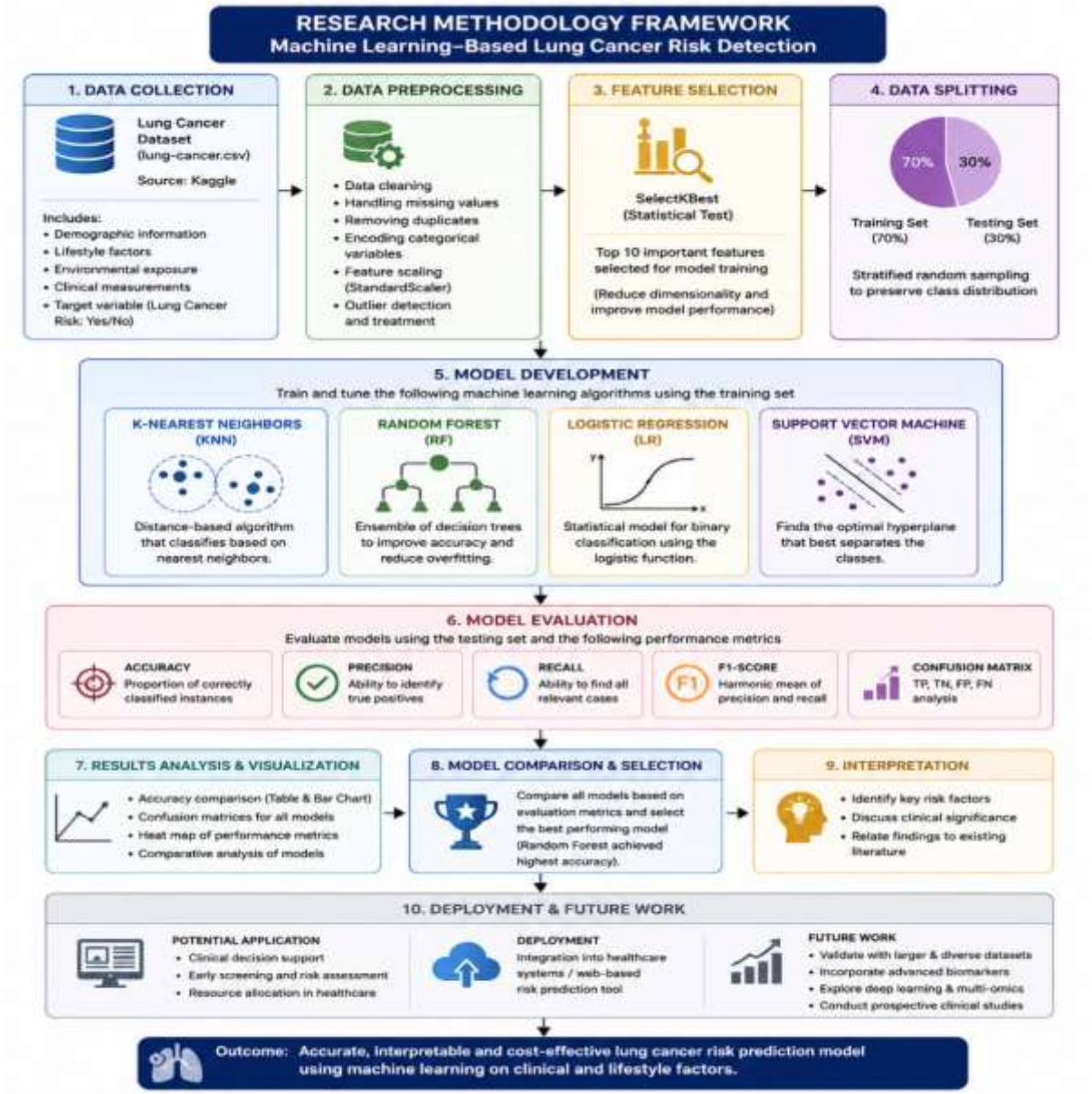
RESEARCH METHODOLOGY FRAMEWORK
Machine Learning–Based Lung Cancer Risk Detection
1. DATA COLLECTION
Lung Cancer Dataset (lung-cancer.csv)
Source: Kaggle
Includes:
• Demographic information
• Lifestyle factors
• Environmental exposure
• Clinical measurements
• Target variable (Lung Cancer Risk: Yes/No)
2. DATA PREPROCESSING
• Data cleaning
• Handling missing values
• Removing duplicates
• Encoding categorical variables
• Feature scaling (StandardScaler)
• Outlier detection and treatment
3. FEATURE SELECTION
SelectKBest (Statistical Test)
Top 10 important features selected for model training
(Reduce dimensionality and improve model performance)
4. DATA SPLITTING
70%
30%
Training Set (70%)
Testing Set (30%)
Stratified random sampling to preserve class distribution
5. MODEL DEVELOPMENT
Train and tune the following machine learning algorithms using the training set
K-NEAREST NEIGHBORS (KNN)
Distance-based algorithm that classifies based on nearest neighbors.
RANDOM FOREST (RF)
Ensemble of decision trees to improve accuracy and reduce overfitting.
LOGISTIC REGRESSION (LR)
Statistical model for binary classification using the logistic function.
SUPPORT VECTOR MACHINE (SVM)
Finds the optimal hyperplane that best separates the classes.
6. MODEL EVALUATION
Evaluate models using the testing set and the following performance metrics
ACCURACY
Proportion of correctly classified instances
PRECISION
Ability to identify true positives
RECALL
Ability to find all relevant cases
F1-SCORE
Harmonic mean of precision and recall
CONFUSION MATRIX
TP, TN, FP, FN analysis
7. RESULTS ANALYSIS & VISUALIZATION
• Accuracy comparison (Table & Bar Chart)
• Confusion matrices for all models
• Heat map of performance metrics
• Comparative analysis of models
8. MODEL COMPARISON & SELECTION
Compare all models based on evaluation metrics and select the best performing model (Random Forest achieved highest accuracy).
9. INTERPRETATION
• Identify key risk factors
• Discuss clinical significance
• Relate findings to existing literature
10. DEPLOYMENT & FUTURE WORK
POTENTIAL APPLICATION
• Clinical decision support
• Early screening and risk assessment
• Resource allocation in healthcare
DEPLOYMENT
Integration into healthcare systems / web-based risk prediction tool
FUTURE WORK
• Validate with larger & diverse datasets
• Incorporate advanced biomarkers
• Explore deep learning & multi-omics
• Conduct prospective clinical studies
Outcome: Accurate, interpretable and cost-effective lung cancer risk prediction model using machine learning on clinical and lifestyle factors.

Figure 3.1. Research methodology framework showing the ten sequential stages followed in this study, from data collection and preprocessing through feature selection, model development, evaluation, and deployment, culminating in an accurate, interpretable, and cost-effective lung cancer risk prediction model.

Source: developed by the authors.

The third part of the study involves the integration of a performance evaluation scheme to assess the efficacy of various performance models and facilitate comparative analysis. The sequential methodology for executing the suggested Machine Learning Classification of Lung Cancer Risk Using Clinical and Lifestyle Risk Factors model is depicted in Figure 3.1

3.3. Dataset Description

The study utilized the publicly available lung_cancer.csv dataset obtained from Kaggle. The dataset contains demographic, lifestyle, environmental, and clinical variables associated with lung cancer risk. Variables include age, gender, smoking history, air pollution exposure, respiratory health indicators, educational status, income level, and related clinical measurements. The target variable identifies whether an individual belongs to a high-risk or low-risk lung cancer category. The dataset is suitable for supervised classification tasks and enables comparative evaluation of machine learning algorithms for risk prediction.

3.4. Environmental Setup

The lung cancer classification model was developed using the robust computational capabilities of the anaconda distribution using the lung_cancer.csv (https://www.kaggle.com/datasets/dhrubangtalukdar/lung-cancer-prediction-dataset.) dataset from Kaggle, an open-source data-sharing platform, to enhance lung cancer detection. A 64-bit Windows operating system, with an Intel(R) Core (TM) i5-8365U CPU @ 1.60 GHz (base) and 1.90 GHz (max turbo), and 16.0 GB (15.8 GB usable) of Random Access Memory (RAM) was used. The programming environment utilized for implementing the program code was Jupyter Notebook, using the Anaconda distribution as the development environment. The application programming interface (API) utilized was the scikit-learn (sklearn) API, together with other Python dependencies such as NumPy for vector operations, pandas for reading files, and Flask for deployment purposes. Matplotlib, a versatile visualization tool designed for machine learning models, is employed to illustrate the graphical behavior of the meticulously implemented models.

3.4.1. Choice of Programming Language and Environment

The choice to employ anaconda distribution as the programming environment for developing the lung cancer detection model was influenced by a combination of factors, each playing a crucial role in enhancing the overall efficiency and efficacy of the model. Anaconda is a popular distribution specifically designed for python and R programming for data computation. Python was chosen as the fundamental programming language for its simplicity and readability, its extensive libraries and frameworks (such as pandas, NumPy, and Flask), cross-platform compatibility, strong community support, and its wide adoption in AI and machine learning applications.

3.5. Dataset Preprocessing

The procedure of data preparation is a crucial element in the execution of any machine learning approach. The main aim of data preparation in this study is to improve the quality and suitability of the data for the specific machine learning task. This study adopts a structured approach to data preparation, encompassing a sequence of phases

designed to efficiently handle the data. The steps involved in this process include data cleaning (checking for missing values), data standardization/normalization, feature selection, and data splitting.

3.5.1. Normalization

Min-Max normalization was applied to transform numerical features into a common scale ranging between 0 and 1. This approach prevents variables with larger numerical ranges from dominating the learning process and improves the convergence and performance of distance-based and gradient-based machine learning algorithms.

3.5.2. Feature Scaling

Feature scaling is a crucial preprocessing step that ensures all input features are on a similar scale. This is particularly important because datasets for Lung cancer risk classification often include diverse features, such as pack-years of smoking, air pollution index, pulmonary function measurements, and demographic information, which may vary significantly in their units and ranges. For example, pack-years may range from 0 to over 60, while pulmonary function measurements such as forced expiratory volume in one second (FEV1) are recorded on an entirely different numerical scale. Without scaling, features with larger magnitudes could dominate the model's learning process, leading to biased or suboptimal results. Feature scaling, particularly standardization, is therefore a vital step in preparing data for lung cancer risk classification models. It ensures that all features contribute equally to the model's learning process, improves algorithm performance, and enhances the interpretability of the results, helping to build more accurate and reliable models for early detection.

3.5.3. Feature Selection

Feature selection is a critical step in machine learning and data analysis. It involves selecting a subset of the most relevant features (variables, predictors) from the original dataset to improve model performance, reduce overfitting, and enhance interpretability. By carefully selecting the most relevant features, the machine learning models were built to be simpler, faster, and more accurate. The choice of feature selection method depends on the dataset, the problem, and the computational resources available. The SelectKBest method with the Analysis of Variance (ANOVA) F-test (f_classif) was used to select the features:

- The f-classify function calculates the F-value for each feature.
- SelectKBest ranks the features based on their F-values and selects the top K features.

3.5.4. Percentage Splitting Technique

The dataset, which represents the lung_cancer.csv, was split into a training set (70%) and a testing set (30%) for evaluating model performance. After experimenting with different split ratios, the best results were achieved at 70:30. This split data was used to train and evaluate the LR, RF, KNN, and SVM models. In data science, the optimal data-splitting ratio depends on the size of the dataset, and there is no universal consensus based on theoretical or empirical research. Although the available dataset in this study is not large, allocating 30% for testing was considered sufficient to provide a reliable evaluation of predictive performance, with the remaining 70% used for training.

3.5.5. Hyperparameter Configuration

The machine learning algorithms were configured using the following parameters:

Logistic Regression

- max_iter=1000
- solver='lbfgs'

Random Forest

- n_estimators=100
- max_depth=None
- random_state=42

KNN

- n_neighbors=5
- metric='euclidean'

SVM

- kernel='rbf'
- C=1.0
- gamma='scale'

3.5.6. Cross-Validation

To improve model reliability and reduce dependence on a single train-test split, 10-fold cross-validation was performed during model development. The dataset was partitioned into ten equal subsets, where nine subsets were used for training and one subset for validation. This process was repeated ten times, and the average performance scores were reported.

3.6. Classification Model

Based on the dataset adopted, this study was framed as a binary classification problem, in which each individual is assigned to one of two risk labels. Classification, in this context, involves the model predicting the correct label, high-risk or low-risk, for lung cancer given a set of input features. Hence, for the identification and classification of lung cancer risk, the study proposed the application of four machine learning models, including LR, RF, KNN, and SVM.

3.6.1. LR

Logistic regression is a statistical method for analyzing datasets in which one or more independent variables determine an outcome. The outcome is measured using a dichotomous variable, meaning there are only two possible outcomes.

3.6.2. RF

Random Forest (RF) combines multiple decision trees to improve the accuracy and robustness of predictions.

3.6.3. KNN

KNN (k-Nearest Neighbors) is a simple, non-parametric, and instance-based machine learning algorithm used for both classification and regression tasks. It works by finding the k closest data points (neighbors) in the feature space to a given input and making predictions based on those neighbors. It is expressed mathematically as shown in Equation 3.1:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (3.1)$$

3.6.4. SVM

SVM is used as a robust classifier for predicting lung cancer risk. It leverages the preprocessed dataset to find the optimal hyperplane that separates the classes. The model's performance is evaluated using various metrics, and its results are visualized using a confusion matrix. SVM is a powerful choice for this task due to its ability to handle high-dimensional data and model non-linear decision boundaries using kernels.

3.7. Performance Evaluation

To evaluate the performance of the model in classifying lung cancer risk, the accuracy, precision, recall and f1-score metrics were adopted, consistent with recommended practices for reporting the performance of classification models (Rainio et al., 2024). The criterion for each of the metrics is calculated according to four main criteria such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) as follow:

True Positive (TP): Indicate instances where the model correctly identifies lung cancer person as high-risk for lung cancer.

False Negative (FN): Indicate instances where the model fails to identify a lung cancer person, instead misclassifying them as non- risk for lung cancer.

False Positive (FP): Indicates the instances where the model incorrectly identifies a low-risk individual as high-risk for lung cancer.

True Negative (TN): Indicates where the model correctly classifies an individual as non-risk for lung cancer.

Precision: Quantifies the accuracy of the model in predicting high-risk for lung cancer. It calculates the proportion of true positive predictions among all positive prediction equation 3.2 depicts the mathematical expression for the precision metrics:

$$Precision = \frac{TP}{TP + FP} \quad (3.2)$$

Recall: Recall, also known as sensitivity or true positive rate, measures the proportion of actual high-risk lung cancer cases that the model correctly identifies. It calculates the proportion of true positive predictions among all actual positive cases. It is mathematically expressed as shown in Equation 3.3:

$$Recall = \frac{TP}{TP+FN} \qquad (3.3)$$

F-score: The F1-score is the harmonic mean of precision and recall, providing a balanced evaluation of the model's performance. It considers both false positive and false negative and is particularly useful imbalanced datasets. It is mathematically expressed in Equation 3.4:

$$F\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (3.4)$$

Accuracy: This is a widely used metric that evaluates the overall performance of the classification model. It measures the proportion of correctly classified instances, including both true positives and true negatives, out of the total number of instances. The accuracy calculation is expressed mathematically in Equation 3.5:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (3.5)$$

4.0. Results and Discussions

4.1. Introduction

This chapter presents the results obtained from the practical exploration and implementation of the proposed models for lung cancer risk classification. As the underlying problem is a classification problem, the study employed classification algorithms, specifically KNN, RF, LR, and SVM, with the results for each presented in tabular and graphical form. The effectiveness of the constructed models was assessed through a thorough validation process, evaluating performance metrics such as accuracy, precision, recall, and F1-score. Prior to presenting the results, data visualization and exploration were conducted to understand the correlations among the dataset features, using graphical visualizations produced with the Seaborn and Matplotlib libraries.

4.2. Discussion of Findings

The experimental results indicate that all four machine learning algorithms achieved strong predictive performance, with classification accuracies exceeding 93%, as summarized in Table 4.1. Among the evaluated models, Random Forest produced the highest accuracy of 95.0%, outperforming Logistic Regression, Support Vector Machine, and K-Nearest Neighbors.

The superior performance of Random Forest may be attributed to its ensemble learning architecture, which combines multiple decision trees to capture complex nonlinear relationships among clinical and lifestyle risk factors. This finding is consistent with previous studies that reported strong predictive performance of tree-based models in lung cancer risk prediction.

The results further suggest that routinely available clinical and lifestyle indicators contain sufficient predictive information for effective lung cancer risk classification. This observation supports previous findings that accessible healthcare data can provide valuable alternatives to expensive molecular diagnostic approaches.

Table 4.1. Accuracy scores achieved by the four machine learning models (Logistic Regression, Random Forest, K-Nearest Neighbors, and Support Vector Machine) when evaluated on the 30% held-out test set, expressed as a percentage of correctly classified instances.

Algorithm	Accuracy (%)
KNN	93.3%
SVM	94.0%
LR	94.2%
RF	95.0%

4.3. Feature Selection

SelectKBest was used to select the top 10 features used to train the models; the selected features are shown in Figure 4.1.

```
0]: # Check the selected features
    selected_features = selector.get_support(indices=True)
    print("\nSelected Features:")
    print(X.columns[selected_features])
```

```
Selected Features:
Index(['smoker', 'smoking_years', 'cigarettes_per_day', 'pack_years',
       'chronic_cough', 'shortness_of_breath', 'oxygen_saturation', 'fev1_x10',
       'crp_level', 'xray_abnormal'],
      dtype='object')
```

Figure 4.1. Top 10 features selected by the SelectKBest algorithm using the Analysis of Variance (ANOVA) F-test, ranked by their statistical relevance to lung cancer risk classification. Source: developed by the authors from the study's model output.

4.4. Data Splitting

This section presents the result of the data-splitting process, in which the dataset was divided in a 70:30 ratio, as shown in Figure 4.2.

Train and Test Data Sizes:
Training data (X_train): 3500 samples
Testing data (X_test): 1500 samples
Training target (y_train): 3500 samples
Testing target (y_test): 1500 samples

Figure 4.2. Proportion of the dataset allocated to model training and testing using a stratified 70:30 split, ensuring that the class distribution of lung cancer risk was preserved in both subsets. Source: developed by the authors from the study's model output.

4.5. Evaluation Metrics Result

In addition to accuracy, the study evaluated each model's performance using the precision, recall, and F1-score metrics, as presented in Table 4.2.

Table 4.2. Precision, recall, and F1-score obtained for each of the four machine learning models on the test set, providing a more detailed view of classification performance beyond overall accuracy.

Algorithm	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
KNN	93.2 %	93.4%	93.2%	93.2%
SVM	94.6%	94.1%	94.2%	94.1%
LR	94.7%	94.2%	94.3%	94.2%
RF	95.0%	95.0%	95.0%	95.0%

4.6. Confusion Matrix

While the confusion matrix is not itself a performance metric, it is an essential diagnostic tool that shows the number of correctly and incorrectly predicted class labels for each model. The confusion matrix results for KNN, SVM, LR, and RF, shown in Figures 4.3, 4.4, 4.5, and 4.6 respectively, indicate strong classification performance in lung cancer risk detection.

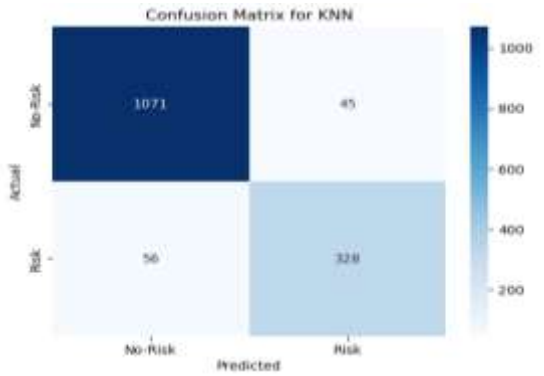

Figure 4.3. Confusion matrix for the K-Nearest Neighbors (KNN) model, showing the counts of true positive, true negative, false positive, and false negative predictions on the test set. Source: developed by the authors from the study's model output.

Figure 4.4. Confusion matrix for the Support Vector Machine (SVM) model, showing the counts of true positive, true negative, false positive, and false negative predictions on the test set. Source: developed by the authors from the study's model output.

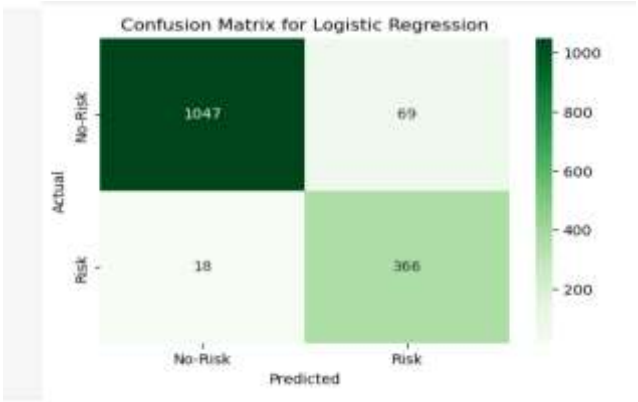

Figure 4.5. Confusion matrix for the Logistic Regression (LR) model, showing the counts of true positive, true negative, false positive, and false negative predictions on the test set. Source: developed by the authors from the study's model output.

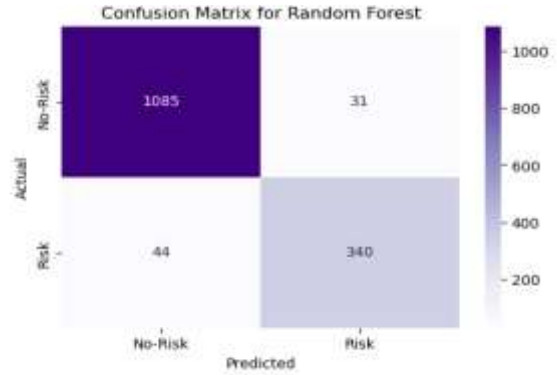

Figure 4.6. Confusion matrix for the Random Forest (RF) model, showing the counts of true positive, true negative, false positive, and false negative predictions on the test set. RF achieved the fewest misclassifications among the four models. Source: developed by the authors from the study's model output.

4.7. Heat Map for the used Models

This section presents a comprehensive heat map of all trained models based on their performance metrics, as shown in Figure 4.7.

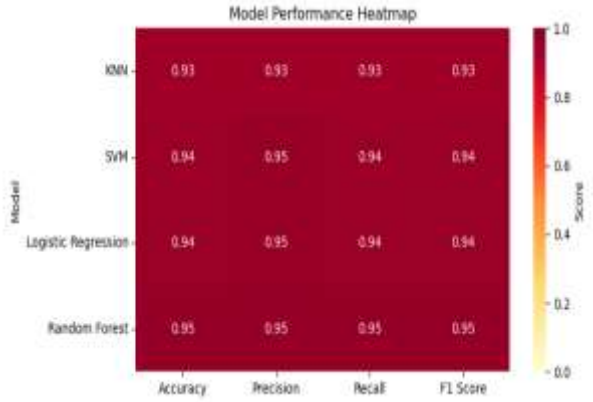

Figure 4.7. Heat map comparing the accuracy, precision, recall, and F1-score of the four machine learning models (KNN, SVM, Logistic Regression, and Random Forest), with darker shading indicating higher scores. Source: developed by the authors from the study's model output.

4.8. Comparison of KNN, SVM, LR and RF Based on Confusion Matrices

The four confusion matrices in Figures 4.3, 4.4, 4.5, and 4.6 correspond to the KNN, SVM, LR, and RF classifiers, respectively. The structure of the matrices differs slightly across models, with RF achieving the highest accuracy of 95.0%, meaning that it made the highest proportion of correct predictions, as shown in Figure 4.8 below.

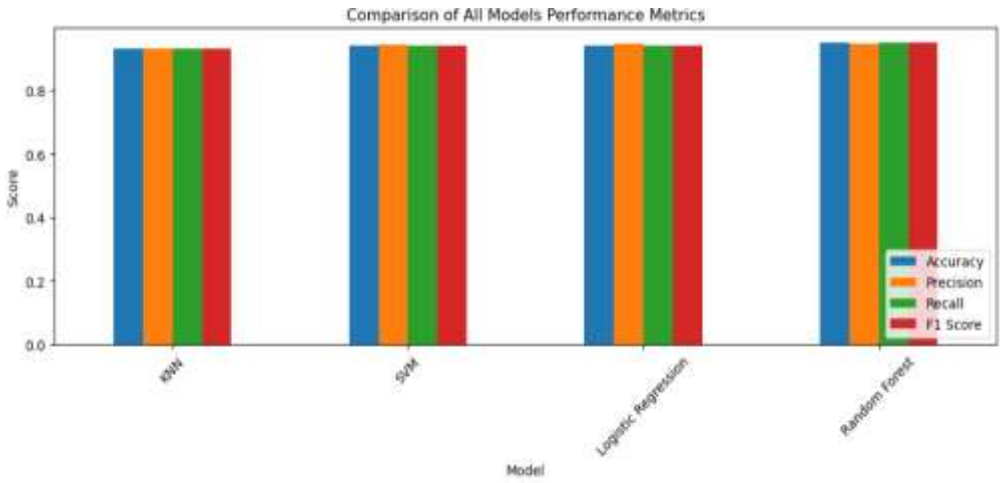

Figure 4.7. Side-by-side comparison of the accuracy, precision, recall, and F1-score achieved by Logistic Regression, Random Forest, K-Nearest Neighbors, and Support Vector Machine, illustrating the overall superiority of Random Forest across all four-evaluation metrics. Source: developed by the authors from the study's model output.

5.0. Conclusion and Future Recommendations

5.1. Summary

This study focused on the development and evaluation of machine learning classification models designed to estimate lung cancer risk using the *lung-cancer.csv* dataset together with related clinical variables. The motivation for the research stems from the substantial global burden of lung cancer, where delayed diagnosis frequently results in poor survival outcomes. Although diagnostic technologies such as low-dose computed tomography (LDCT) have demonstrated effectiveness in detecting early disease, their availability remains limited in many healthcare environments, particularly in resource-constrained regions. This limitation highlights the importance of exploring affordable and scalable alternatives.

To address this need, the research utilized a publicly available dataset obtained from Kaggle, which includes demographic attributes, lifestyle characteristics, environmental exposure indicators, and clinical measurements relevant to lung cancer risk. A systematic analytical process was applied that included data preprocessing, normalization of variables through feature scaling, and the selection of the most informative predictors using the SelectKBest feature selection method.

Four machine learning algorithms Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) were implemented and trained using 70% of the dataset, while the remaining 30% was reserved for performance evaluation. Model effectiveness was assessed using widely accepted classification metrics, including accuracy, precision, recall, and F1-score.

The experimental findings indicated that all four algorithms performed strongly, each achieving classification accuracies above 93%. Among them, the Random Forest model demonstrated the best overall performance, producing an accuracy rate of 95.0% and equally strong values for precision, recall, and F1-score. These results suggest that ensemble-based learning approaches are particularly effective for identifying individuals who may be at increased risk of developing lung cancer.

Overall, the study demonstrates that machine learning techniques can successfully utilize routinely available clinical and lifestyle data to generate reliable predictive models. Such models offer considerable promise as decision-support tools for lung cancer risk assessment, providing answers to the central research questions concerning predictive accuracy and the relevance of selected risk factors.

5.2. Study Limitations

Despite promising results, several limitations should be acknowledged.

1. The study relied on a publicly available Kaggle dataset rather than real-world clinical records.
2. External validation using independent datasets was not performed.
3. The evaluation utilized a limited number of machine learning algorithms.
4. The dataset primarily contained clinical and lifestyle risk factors rather than advanced molecular biomarkers.
5. Subgroup analysis across demographic categories was not conducted.
6. These limitations should be addressed in future investigations to improve generalizability and clinical applicability.

5.3. Conclusion

This study demonstrates that machine learning algorithms can effectively utilize clinical and lifestyle risk factors for lung cancer risk classification. Among the evaluated models, Random Forest achieved the highest predictive performance, suggesting that ensemble learning approaches are particularly effective for capturing complex interactions among risk variables. The findings further indicate that affordable and routinely collected healthcare data may provide a practical foundation for supporting early lung cancer risk assessment in resource-constrained environments.

1. Strong Predictive Performance

The machine learning models developed in this study achieved high predictive performance, with the Random Forest algorithm delivering the most accurate results at 95.0%. This level of accuracy demonstrates that machine learning approaches are capable of reliably distinguishing between individuals with elevated lung cancer risk and those with lower risk based on the selected predictive variables.

2. Value of Easily Obtainable Clinical Data

Another significant outcome of the study is the confirmation that commonly available medical information including routine laboratory measurements, smoking exposure history, respiratory health indicators, and

demographic characteristics can serve as powerful predictors in machine learning models. Because these variables are relatively inexpensive and widely accessible, they offer a practical foundation for developing screening tools that do not rely on expensive imaging technologies or specialized laboratory tests.

3. Advantages of Ensemble Learning Techniques

Comparative evaluation of the four algorithms revealed that the Random Forest classifier performed better than Logistic Regression, Support Vector Machine, and K-Nearest Neighbors. The improved performance can be attributed to the ensemble structure of Random Forest, which aggregates multiple decision trees to capture complex and nonlinear relationships within the dataset. This capability enhances both model stability and predictive accuracy.

4. Addressing the Research Questions

The study also successfully responded to the key research questions guiding the investigation:

a) Model Accuracy

The classification models, particularly Random Forest, demonstrated strong predictive accuracy with high precision and recall values, indicating reliable identification of individuals at risk.

b) Important Predictive Factors

The feature selection process identified ten key predictors contributing most significantly to the classification process. Among the most influential variables were *pack_years*, *air_pollution_index*, *fev1_x10*, and *crp_level*, confirming their relevance in lung cancer risk prediction.

c) Performance Across subgroups

Although subgroup-specific analysis was not explicitly conducted, the strong overall model performance suggests that the model functions effectively across the diversity present within the dataset, including variations in age, gender, and smoking status. Future research could further investigate model behavior within specific demographic groups.

d) Comparison with Conventional Risk Assessment

Unlike traditional approaches that often rely on single-variable criteria, the machine learning model integrates multiple risk factors simultaneously to generate a comprehensive risk estimate. This multi-factor analytical capability represents a meaningful improvement over conventional screening strategies.

In summary, the study provides strong evidence that machine learning can be effectively integrated into healthcare systems as a non-invasive, affordable, and scalable approach to early lung cancer risk identification.

5.4. Recommendations

Based on the outcomes of the research as well as the limitations identified during the study, several recommendations are proposed for both practical implementation and future research directions.

5.4.1. Potential Application Areas

The developed machine learning model has several promising real-world applications that could contribute to improved lung cancer detection strategies.

I. Clinical Screening Support in Primary Care

One potential application involves integrating the model into electronic health record (EHR) systems used in primary healthcare settings. By analyzing routine blood test results and patient history, the model could automatically identify individuals with elevated risk profiles. Healthcare providers could then recommend additional diagnostic procedures, such as LDCT scans, for these patients, potentially improving early detection rates.

II. Population-Level Risk Assessment for Public Health

Public health organizations could also apply this predictive model to large population datasets to identify communities or demographic groups with higher predicted lung cancer risk. Such insights would enable targeted allocation of screening programs and preventive interventions, including smoking cessation initiatives and environmental risk mitigation strategies.

III. Clinical Triage in Resource-Limited Settings

In healthcare environments where advanced imaging technologies are unavailable or financially inaccessible, the model could function as an initial triage tool. Patients identified as high risk could be prioritized for further medical evaluation or referral to specialized healthcare facilities. This approach would help ensure that limited healthcare resources are directed toward individuals who need them most.

5.4.2. Directions for Future Research

While the current study provides promising results, several avenues remain for further investigation to strengthen and expand upon these findings.

I. Validation with Larger and More Diverse Datasets

A critical next step involves evaluating the model using independent datasets from multiple geographic regions and healthcare systems. External validation would help determine whether the model maintains its predictive performance across diverse populations and clinical contexts.

II. Prospective Clinical Evaluation

Following successful validation, the model should be assessed through prospective clinical studies. Such trials would allow researchers to examine the model's real-world impact on early diagnosis rates, treatment outcomes, and overall healthcare costs, as well as its ability to integrate into routine clinical practice.

III. Implementation of Explainable Artificial Intelligence

To increase transparency and encourage clinical adoption, future work should incorporate advanced explainable AI methods. Techniques such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic

Explanations (LIME) can provide detailed insights into how individual variables influence specific predictions, thereby enhancing clinician trust in machine learning systems.

IV. Development of Prognostic Prediction Models

The present research focused primarily on diagnostic classification that is, identifying individuals who may already be at risk of lung cancer. Future studies could expand this work by developing prognostic models capable of predicting disease progression, treatment response, or survival outcomes. Such models would contribute significantly to personalized patient management and clinical decision-making.

V. Integration into Real-Time Clinical Decision Support Systems

Future work should explore embedding the validated model into electronic health record (EHR) systems or web-based clinical decision support tools, enabling clinicians to obtain real-time risk estimates during routine consultations and thereby supporting timelier referral for further diagnostic investigation.

VI. Cost-Effectiveness and Health Economics Analysis

Subsequent research should evaluate the cost-effectiveness of deploying the proposed model relative to conventional screening approaches, particularly in resource-constrained healthcare settings, to provide policymakers with the economic evidence needed to support wider adoption.

Declarations

Source of Funding

The authors received no specific funding for this research.

Competing Interests Statement

The authors declare that there are no competing interests regarding the publication of this manuscript.

Consent for Publication

The authors declare that they consented to the publication of this article.

Authors' Contributions

Author 1: Conceptualization, methodology, software development, data analysis, writing of original draft. Author 2: Validation, supervision, review, and editing. Author 3: Investigation, visualization, proofreading, and manuscript revision.

Informed Consent

Not Applicable.

Availability of Data and Materials

The dataset utilized in this study (lung_cancer.csv) is publicly available on Kaggle, an open-access data-sharing platform (Lung Cancer Risk Prediction Dataset).

Institutional Review Board Statement

Not Applicable.

Ethical Approval

Not Applicable. The study utilized anonymized secondary data obtained from publicly accessible repositories.

Acknowledgement

The authors acknowledge Kaggle and the dataset contributor for providing public access to the lung_cancer.csv dataset and appreciate the support provided by their respective institutions during the conduct of this study.

Declaration of Artificial Intelligence

Artificial Intelligence tools were utilized solely for language enhancement, grammar checking, and editorial support during manuscript preparation. All scientific interpretations, analyses, conclusions, and final content were independently reviewed and validated by the authors.

References

Ankolekar, A., Boie, S., Abdollahyan, M., Gadaleta, E., Hasheminasab, S.A., Yang, G., et al. (2025). Advancing breast, lung and prostate cancer research with federated learning: A systematic review. NPJ Digital Medicine, 8: 314. https://doi.org/10.1038/s41746-025-01173-2.

Arguís, M., Onoiu, A.I., Castañé, H., Camps, J., Arenas, M., & Joven, J. (2024). Combining metabolomics and machine learning to identify diagnostic and prognostic biomarkers in patients with non-small cell lung cancer pre- and post-radiation therapy. Biomolecules, 14(8): 898. https://doi.org/10.3390/biom14080898.

Chen, S., Han, F., Zhao, X., Zhou, H., Tang, H., Gao, Q., & Chen, J. (2023). Integrated models of blood protein and metabolite enhance the accuracy of non-small cell lung cancer detection. Biomarker Research, 11: 68. https://doi.org/10.1186/s40364-023-00512-4.

Davies, M.E., Chaimowitz, M.I., Metcalf, M., Kelaini, S., Borras, E., Prensner, J.R., & Cramer, D.W. (2023). Plasma protein biomarkers for early prediction of lung cancer. eBioMedicine, 91: 104567. https://doi.org/10.1016/j.ebiom.2023.104567.

Dritsas, E., & Trigka, M. (2022). Lung cancer risk prediction with machine learning models. Big Data and Cognitive Computing, 6(4): 139. https://doi.org/10.3390/bdcc6040139.

Fehlmann, T., Kahraman, M., Ludwig, N., Backes, C., Fehlmann, T., et al. (2020). Evaluating the use of circulating microRNA profiles for lung cancer detection in symptomatic patients. JAMA Oncology, 6(8): 1217–1223. https://doi.org/10.1001/jamaoncol.2020.1457.

Flyckt, R.N.H., Sjödsholm, L., Henriksen, M.H.B., Unnikrishnan, A., Petersen, S.S., Kaltoft, M.S., et al. (2024). Pulmonologist-level lung cancer detection based on standard blood test results and smoking status using an

explainable machine learning approach. Scientific Reports, 14: 30630. https://doi.org/10.1038/s41598-024-30630-0.

Gandhi, Z., Gurram, P., Amgai, B., Lekkala, S.P., Lokhandwala, A., Manne, S., et al. (2023). Artificial intelligence and lung cancer: Impact on improving patient outcomes. Cancers, 15(21): 5236. https://doi.org/10.3390/cancers15215236.

Heitzer, E., White, R., Risch, L., Kobold, A.C.B., Lehr, M., Heidrich, V., et al. (2023). Bridging biological cfDNA features and machine learning approaches: A clinician's perspective on liquid biopsy. Trends in Genetics, 39(4): 371–377. https://doi.org/10.1016/j.tig.2022.12.006.

Hu, S., Tao, J., Peng, M., Ye, Z., Chen, Z., Chen, Z., & Ni, B. (2023). Accurate detection of early-stage lung cancer using a panel of circulating cell-free DNA methylation biomarkers. Biomarker Research, 11: 45. https://doi.org/10.1186/s40364-023-00480-9.

International Agency for Research on Cancer (2023). Global Cancer Observatory: Cancer today. https://gco.iarc.fr/today.

Ji, Z., Wang, F., Mu, C., Yin, T., Long, T., Shen, Y., & Cheng, L. (2023). Serum untargeted metabolomics reveal metabolic alterations of non-small cell lung cancer and refine disease detection. Cancer Science, 114(3): 919–933. https://doi.org/10.1111/cas.15641.

Karymsakova, I., Kozhakhmetova, D., Bekenova, D., Abdullah, L.N., Zolotov, A., Shyrynkhanova, D., Kydyralina, L., & Bugubayeva, A. (2026). Application of machine learning methods for predicting susceptibility to lung cancer and identifying predictors influencing the risk of lung cancer development. Computers, 15(6): 365. https://doi.org/10.3390/computers15060365.

Kim, D.H., Park, H., Choi, Y.J., Im, K., Lee, C.W., Kim, D.S., & Rho, J.K. (2023). Identification of exosomal microRNA panel as diagnostic and prognostic biomarker for small cell lung cancer. Biomarker Research, 11: 80. https://doi.org/10.1186/s40364-023-00540-0.

Kwon, H.J., Park, U.H., Gu, C.J., Park, D.B., Lee, Y.G., Lee, I.K., & Lee, M.S. (2023). Enhancing lung cancer classification through integration of liquid biopsy multi-omics data with machine learning techniques. Cancers, 15(18): 4556. https://doi.org/10.3390/cancers15184556.

Li, B., Cui, L., Yang, C., Li, W., Liu, J., Li, J., & Tang, H. (2025). Machine learning-derived peripheral blood transcriptomic biomarkers for early lung cancer diagnosis. BioFactors. Advance Online Publication. https://doi.org/10.1002/biof.1934.

Liu, Z., Huang, H., Ren, J., Song, T., Ni, Y., Mao, S., & Tang, H. (2024). Plasma exosomes contain protein biomarkers valuable for the diagnosis of lung cancer. Discover Oncology, 15: 194. https://doi.org/10.1007/s12672-024-00934-3.

Maharjan, N., Thapa, N., & Tu, J. (2020). Blood-based biomarkers for early diagnosis of lung cancer: A review article. Journal of Nepal Medical Association, 58(227): 519–524. https://doi.org/10.31729/jnma.5038.

Mathios, D., Bettegowda, C., Szantai-Kis, D.M., Sammut, S.J., Dietz, S., et al. (2021). Detection and characterization of lung cancer using cell-free DNA fragmentomes. Nature Communications, 12: 5065. https://doi.org/10.1038/s41467-021-24994-8.

Poh, C.F., Isin, M.N., Mazlan, A.M., Baharudin, R., & Wahid, N.A.A. (2023). Development of a microRNA-based model for lung cancer detection. Cancers, 15(4): 1112. https://doi.org/10.3390/cancers15041112.

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. Scientific Reports, 14: 6086. https://doi.org/10.1038/s41598-024-56734-0.

Røe, O.D., Du, Y., Li, R., Zhang, Y., & Ivanova, O. (2024). Promising microRNAs in pre-diagnostic serum associated with lung cancer up to eight years before diagnosis: A HUNT study. Journal of Cancer Research and Clinical Oncology. Advance Online Publication. https://doi.org/10.1007/s00432-024-05411-5.

Schwartz, L., Matania, N., Levi, M., Lazebnik, T., Kushnir, S., Yosef, N., Hoogi, A., & Shlomi, D. (2025). A blood test-based machine learning model for predicting lung cancer risk. Frontiers in Medicine, 12: 1577451. https://doi.org/10.3389/fmed.2025.1577451.

Sharma, R. (2022). Mapping of global, regional and national incidence, mortality and mortality-to-incidence ratio of lung cancer in 2020 and 2050. International Journal of Clinical Oncology, 27(4): 665–675. https://doi.org/10.1007/s10147-021-02108-2.

Thalambedu, N., Balla, M., Sivasubramanian, B.P., Sadaram, P., Prathiba, K., Kruparani, N., et al. (2025). Integrating artificial intelligence with circulating tumor DNA for non-small cell lung cancer: Opportunities, challenges, and future directions. Frontiers in Medicine, 12: 1612376. https://doi.org/10.3389/fmed.2025.1612376.

Tsai, Y.T., Schlom, J., & Donahue, R.N. (2024). Blood-based biomarkers in patients with non-small cell lung cancer treated with immune checkpoint blockade. Journal of Experimental & Clinical Cancer Research, 43: 82. https://doi.org/10.1186/s13046-024-02962-3.

World Health Organization (2023). Lung cancer fact sheet. https://www.who.int/news-room/fact-sheets/detail/lung-cancer.

Wu, J., Zan, X., Gao, L., Zhao, J., Fan, J., Shi, H., Wan, Y., Yu, E., Li, S., & Xie, X. (2019). A machine learning method for identifying lung cancer based on routine blood indices: Qualitative feasibility study. JMIR Medical Informatics, 7(3): e13476. https://doi.org/10.2196/13476.